



ANALISIS KELAS LATEN (*LATEN CLASS ANALYSIS*) UNTUK PENGELOMPOKAN DATA KATEGORIK

Timbul Pardede
FMIPA-Universitas Terbuka

timbul@mail.ut.ac.id

Dalam analisis kelompok pada bidang sosial, seringkali para peneliti menggunakan instrumen sebagai sumber informasi data yang akan dianalisis. Pada banyak kasus, kuesioner digunakan untuk mengukur suatu variabel yang tidak dapat diukur secara langsung karena objek yang diamati tidak memiliki nilai kuantitatif. Untuk dapat mengukur variabel yang tidak dapat diukur secara langsung, digunakan variabel indikator (manifes) yang memiliki tipe data kategorik. Dengan informasi yang diperoleh dari variabel indikator dibentuk sebuah variabel laten. Alat statistik yang dapat digunakan untuk mengelompokkan variabel laten adalah analisis Kelas laten (*Latent Class Analysis*). Pendugaan parameter dalam analisis analisis Kelas laten digunakan metode *Maximum Likelihood Estimation* (MLE) dengan iterasi algoritma *Expectation-Maximum* (EM). Dengan pendekatan *Bayesian Information Criterion* (BIC) dan *Akaike Information Criterion* (AIC) diperoleh model terbaik. Untuk melihat kecocokan model digunakan Statistik Pearson Chi kuadrat (χ^2) dan Statistik ratio *likelihood* (G^2). Penelitian ini bertujuan untuk mengelompokkan data kategorik dengan menggunakan analisis kelas laten. Penelitian ini menggunakan dua data sekunder, *Diagnoses of carcinoma* (karsinoma), data sampel yang terdiri dari 7 dikotomis tentang ada atau tidaknya karsinoma pada servik uterus dan *General Social Survey 1982*, data sampel tentang sikap warga Amerika Serikat terhadap survei sosial. Dengan menggunakan paket program *poLCA* versi 1.4 pada program R versi 3.0.2, hasil penelitian menunjukkan bahwa untuk data karsinoma, model terbaik terdapat pada tiga model kelas laten, yaitu kelompok yang secara konsisten dinilai positif ada penyakit karsinoma diwakili 44,47%, secara konsisten dinilai negatif terkena penyakit karsinoma diwakili 37,36% dari populasi, dan 18,17 % dari populasi dinilai meragukan. Untuk data survei sosial, model terbaik terdapat pada tiga model kelas laten, yakni tipe responden ideal (62,1%), optimis (20,7%), ragu-ragu (17,2%).

Kata Kunci: Variabel laten, Variabel Manifes, Algoritma *EM*, *BIC*, *AIC*, Kriteria kecocokan model.

PENDAHULUAN

Analisis kelompok (*cluster analysis*) merupakan salah satu teknik dalam analisis statistik multivariat yang bertujuan untuk mengelompokkan suatu objek pengamatan menjadi beberapa kelompok objek amatan berdasarkan karakteristik variabel-variabel yang dimiliki, sedemikian sehingga objek-objek amatan yang terletak dalam kelompok yang sama cenderung mempunyai sifat yang relatif lebih homogen dibandingkan dengan objek-objek amatan pada kelompok yang berbeda (Johnson & Wichern, 2007). Pengelompokan objek-objek tersebut dilakukan berdasarkan suatu ukuran kemiripan atau ketidakmiripan. Semakin tinggi kemiripan dua objek pengamatan maka semakin tinggi peluang untuk dikelompokkan dalam suatu kelompok, sebaliknya semakin tinggi ketidakmiripannya maka semakin rendah peluang untuk dikelompokkan dalam suatu kelompok.

Dalam penelitian bidang sosial, seringkali para peneliti menggunakan instrumen penelitian dalam bentuk kuesioner yang berupa data kategorik sebagai sumber data yang akan dianalisis. Melalui kuesioner akan diperoleh informasi dari responden. Pada banyak kasus, Kuesioner digunakan untuk mengukur suatu variabel yang tidak dapat

diukur secara langsung karena objek yang diamati tidak memiliki nilai kuantitatif. Variabel seperti ini disebut dengan variabel laten. Untuk dapat mengukur variabel yang tidak dapat diukur secara langsung (variabel laten), diperlukan suatu variabel indikator. Biasanya variabel indikator berbentuk item-item pertanyaan dalam data kategorik, yang biasa disebut dengan variabel manifes. Dengan menggunakan informasi yang diperoleh dari variabel indikator (manifes) dibentuk sebuah variabel laten.

Salah satu analisis kelompok yang dapat digunakan untuk menganalisis data kategorik ataupun data campuran (data kontinu dan kategorik) adalah analisis kelas laten (*Latent Class Analysis*). Analisis kelas laten merupakan alat yang berguna untuk mengidentifikasi kelompok dalam data kategorik maupun data campuran multivariat. Menurut Magidson dan Vermunt (2001), analisis kelas laten merupakan teknik pemodelan statistik yang banyak digunakan dalam penelitian ilmu perilaku dan sosial untuk mengidentifikasi keanggotaan kelompok yang tidak teramati (*unobservable*) antara obyek dengan variabel yang diamati. Masing-masing variabel pengamatan dalam analisis kelas laten diasumsikan memiliki distribusi bersyarat sesuai dengan skala datanya (nominal, ordinal, interval, dan ratio). Analisis kelas laten menggunakan model peluang untuk memaksimalkan kecocokan model terhadap data. Model tersebut dapat mengidentifikasi pola dari beberapa variabel tak bebas (multi-variabel). Selanjutnya setiap responden dimasukkan ke dalam setiap kelompok yang terbentuk berdasarkan peluang terbesarnya.

Berdasarkan paparan diatas, tujuan penelitian ini adalah untuk melakukan analisis kelompok terhadap data bertipe kategorik (nominal, ordinal) dengan menggunakan analisis laten kelas.

Analisis Kelas Laten

Analisis kelas laten diperkenalkan pada tahun 1950 oleh Lazarsfeld (dalam Linzer dan Jerey, 2011) dengan sebutan analisis struktur laten. Pada tahun 1968, Lazarsfeld dan Henry (dalam Magidson dan Vermunt, 2001) model kelas laten sebagai suatu alat untuk membentuk kelompok data berdasarkan variabel dikotomis yang diamati. Ide yang mendasari analisis kelas laten ini adalah beberapa parameter dari model statistik yang dialihkan berbeda pada subkelompok-subkelompok yang tidak teramati (laten). Subkelompok-subkelompok ini mengasumsikan variabel laten adalah data kategorik. Model kelas laten tidak tergantung pada asumsi pemodelan yang sering dilanggar dalam praktek (misalnya: hubungan linear, distribusi normal, homogenitas). Analisis kelas laten mendefinisikan kelas laten dengan kriteria kondisi yang independen, artinya bahwa masing-masing variabel secara statistik independen dengan variabel lain dalam setiap kelas laten. Moustaki dan Papageorgiou (2004)

menjelaskan tujuan dari analisa kelas laten adalah untuk mengidentifikasi jumlah kelas yang diperlukan untuk menjelaskan hubungan antara variabel-variabel yang diamati dan untuk mengalokasikan responden masuk dalam kelas laten. Dean N., Raftery, A, R. (2010) menyatakan bahwa kelas laten dapat ditampilkan sebagai kasus khusus dari pengelompokan berdasarkan model (*model-based clustering*) untuk data diskrit multivariat. Pengelompokan berdasarkan model mengasumsikan bahwa setiap pengamatan menjadi satu kelompok dari sejumlah kelompok atau subpopulasi.

Pada analisis kelas laten, Collins dan Lanza (2010) menyatakan bahwa model kelas laten haruslah saling bebas (*local independence*), yang artinya setiap variabel yang berada pada satu kelompok/kelas laten harus saling bebas. Asumsi kebebasan lokal ini hanya berlaku untuk variabel laten saja, bukan setiap variabel yang diamati harus independen. Apabila asumsi ini tidak dipenuhi akan jauh lebih rumit proses analisisnya, karena objek amatan harus dikondisikan tidak hanya pada keanggotaan kelas laten, tetapi juga yang lainnya.

Misalnya pengamatan J merupakan variabel kategorik *polytomous* (variabel manifes), setiap konstanta K_j . Misalkan set data pengamatan (Y_1, Y_2, \dots, Y_J) dengan $Y_j = (Y_{ijk}; j = 1, \dots, J; k = 1, \dots, K_j)$. Dinotasikan Y_{ijk} adalah nilai teramati dari variabel manifes J sedemikian sehingga

$$Y_{ijk} = \begin{cases} 1 & \text{jika individu } i \text{ pada ke-} k \text{ merespon variabel manifes ke-} j \\ 0 & \text{untuk lainnya} \end{cases}$$

Model kelas laten mendekati distribusi gabungan pengamatan dari variabel manifes. Misalkan π_{jrk} merupakan peluang kelas bersyarat yang diamati dalam kelas $r = 1, \dots, R$ yang menghasilkan *outcome* k pada variabel ke- j . R merupakan penduga

awal yang didasarkan pada kecocokan model sehingga $\sum_{k=1}^{K_j} \pi_{jrk} = 1$.

Peluang individu i dalam kelas r dengan variabel manifes Y dengan asumsi saling bebas (*local independence*) adalah:

$$f(Y_i; \pi_r) = \prod_{j=1}^J \prod_{k=1}^{K_j} (\pi_{jrk})^{Y_{ijk}} \quad (1)$$

Fungsi kepekatan peluang di semua kelas laten adalah:

$$P(Y_i | \pi, p) = \sum_{r=1}^R p_r \prod_{j=1}^J \prod_{k=1}^{K_j} (\pi_{jrk})^{Y_{ijk}} \quad (2)$$

dengan p_r adalah peluang awal laten kelas pada kelompok ke r dan π_{jrk} adalah parameter vektor pada kelas laten yang akan diduga. Peluang posterior yang setiap

individu masuk dalam setiap kelas dan tergantung pada nilai variabel manifes yang diamati dihitung dengan menggunakan formula Bayes, yaitu:

$$P(r|Y_i) = \frac{\hat{p}_r f(Y_i; \hat{\pi}_r)}{\sum_{q=1}^R \hat{p}_q f(Y_i; \hat{\pi}_q)} \quad (3)$$

dengan \hat{p}_r adalah penduga p_r dan $\hat{\pi}_{jrk}$ adalah penduga π_{jrk} . Jumlah parameter yang akan diduga adalah $(R-1) + R \sum_j (K_j - 1)$. Jumlah parameter ini akan bertambah besar dengan bertambahnya R, J , dan K_j . Jika jumlah parameter ini melebihi dari jumlah pengamatan atau lebih sedikit satu dari jumlah sel tabel klasifikasi silang variabel manifes maka analisis kelas laten tidak dapat digunakan.

Pendugaan parameter kelas laten pada persamaan (2) dilakukan dengan menggunakan metode *Maximum likelihood Estimation* (MLE) yang dimaksimalkan dengan proses iterasi algoritma *Expectation-Maximum* (EM). Algoritma EM adalah sebuah prosedur iterasi untuk menghitung pendugaan maksimum *likelihood* pada parameter model yang mengandung data tidak teramati. Tiap iterasi dalam algoritma EM terdiri dari dua tahap yaitu tahap E untuk pendugaan dan tahap M untuk pemaksimuman. Dalam tahap E, data hilang diduga berdasarkan data yang diamati. Untuk mendapatkan pendugaan parameter, maka digunakan penduga yang dimaksimumkan berdasarkan fungsi *likelihood*. Fungsi *likelihood* didefinisikan sebagai:

$$L = \sum_{r=1}^R p_r \prod_{j=1}^J \prod_{k=1}^{K_j} (\pi_{jrk})^{Y_{ijk}} \quad (4)$$

Fungsi *log-likelihood*:

$$\log L = \sum_{i=1}^N \ln \left(\sum_{r=1}^R p_r \prod_{j=1}^J \prod_{k=1}^{K_j} (\pi_{jrk})^{Y_{ijk}} \right) \quad (5)$$

Fungsi *log-likelihood* pada persamaan (5) dimaksimumkan dengan kendala $\sum_{r=1}^R p_r = 1$.

Misalkan $z = (z_1, \dots, z_g)$ dengan $z_r = (z_{1r}, \dots, z_{nr})$ dan

$$z_{ir} = \begin{cases} 1, & x_i \in \text{kelompok } r \\ 0, & \text{lainnya.} \end{cases} \quad ; i = 1, \dots, n, \quad r = 1, \dots, R$$

Dengan asumsi z bebas dan identik, maka fungsi kemungkinan data lengkap (*complete-data likelihood*) adalah :

$$\log L = \sum_{i=1}^N \sum_{r=1}^g z_{ik} \ln \left(p_r \prod_{j=1}^J \prod_{k=1}^{K_j} (\pi_{jrk})^{Y_{ijk}} \right) \quad (6)$$

Misalkan pemilihan nilai awal untuk \hat{p}_r dan $\hat{\pi}_{jrk}$ adalah \hat{p}_r^{old} dan $\hat{\pi}_{jrk}^{old}$, maka

- Tahap E: substitusi \hat{p}_r^{old} dan $\hat{\pi}_{jrk}^{old}$ pada persamaan (3), diperoleh peluang bersyarat:

$$\hat{P}(r|Y_i) = \frac{\hat{p}_r f(Y_i; \hat{\pi}_r)}{\sum_{q=1}^R \hat{p}_q f(Y_i; \hat{\pi}_q)} \quad (7)$$

- Tahap M: perbaikan pendugaan parameter dengan memaksimumkan fungsi *log-likelihood*, diperoleh:

$$\hat{p}_r^{new} = \frac{1}{N} \sum_{i=1}^N \hat{P}(r|Y_i) \text{ sebagai peluang awal baru} \quad (8)$$

$$\hat{\pi}_{jr}^{new} = \frac{\sum_{i=1}^N Y_{ij} \hat{P}(r|Y_i)}{\sum_{i=1}^N \hat{P}(r|Y_i)} \text{ sebagai peluang outcome bersyarat} \quad (9)$$

dengan $\hat{\pi}_{jr}^{new}$ merupakan panjang vektor K_j dari peluang *outcome* bersyarat pada kelas r untuk variabel manifes ke- j dan Y_{ij} merupakan matriks $N \times K_j$ matriks hasil pengamatan Y_{ijk} . Demikian proses iterasi algoritma *EM* ini dilakukan hingga kriteria kekonvergenan dipenuhi.

Pemilihan dan Kriteria Kecocokan Model

Collins dan Lanza (2010) mengemukakan pemilihan dan kriteria kecocokan model kelas laten perlu diperhatikan prinsip *parsimony*, yaitu kesederhanaan model. Model yang baik adalah banyak parameter yang diduga lebih sedikit dan adanya keseimbangan antara kelebihan atau kekurangan kecocokan model. Ukuran *parsimony* yang sering digunakan adalah *Bayesian Information Criterion* (BIC) dan *Akaike Information Criterion* (AIC) yang didefinisikan sebagai berikut:

$$AIC = -2\Lambda + 2\Phi$$

dan

$$BIC = -2\Lambda + \Phi \ln N$$

Keterangan:

Λ : nilai maksimum *log-likelihood*

Φ : jumlah parameter model
 N : banyak objek pengamatan.

Nilai AIC dan BIC mencerminkan peningkatan nilai jumlah kuadrat residual dan jumlah parameter model yang digunakan. Nilai AIC dan BIC terkecil menunjukkan bahwa model yang dihasilkan lebih dapat menjelaskan variasi data. Hal ini berarti nilai AIC dan BIC terkecil mengindikasikan bahwa model tersebut adalah model kelas laten yang paling layak. Menurut Lin dan Dayton (dalam Linzer dan Lewis, 2011), nilai BIC lebih sering digunakan untuk menentukan model kelas laten terbaik karena kesederhanaan model.

Kriteria kecocokan model mengacu pada apakah model kelas laten dapat menjelaskan data dengan cukup baik. Ada dua statistik uji yang dapat digunakan untuk menguji kecocokan model, yaitu: Statistik Pearson Chi kuadrat (χ^2) dan Statistik *ratio likelihood* (G^2).

Misalkan ($j = 1, \dots, J$) adalah variabel manifes dan setiap variabel manifes j mempunyai $r_j = 1, \dots, R_j$ kategori. Tabel kontingensi yang dibentuk dari tabulasi silang variabel manifes j mempunyai sel ($c = 1, \dots, C$) dengan $W = \prod_{j=1}^J R_j$. Misalkan Q_c adalah frekuensi sel c dan \hat{Q}_c adalah frekwensi harapan pada sel c yang diformulasikan dengan: $\hat{Q}_c = N \sum_{r=1}^R \hat{p}_r \prod_{j=1}^J \hat{\pi}_{jr_k}$

Kedua statistik uji:

1. Statistik Pearson Chi kuadrat (χ^2)

$$\chi^2 = \sum_c \frac{(q - Q)^2}{\hat{Q}}$$

2. Statistik *ratio likelihood* (G^2)

$$G^2 = 2 \sum_c q \ln \left(\frac{q}{\hat{Q}} \right)$$

Statistik χ^2 mendekati distribusi chi kuadrat dengan derajat bebas $db = W - P - 1$, dengan W adalah banyak sel dan P adalah banyak parameter yang diduga.

METODOLOGI

Sumber data dalam penelitian ini menggunakan dua data sekunder. Data sekunder pertama adalah data *Diagnoses of carcinoma* (diagnosa karsinoma) yang diperoleh dari paket program polCA versi 1.4 pada program R versi 3.02 dan Agresti

(2002). Data diagnosa karsinoma ini merupakan data sampel yang terdiri dari 7 dikotomis (sebagai variabel manifes) yang digunakan untuk mengetahui ada atau tidaknya karsinoma pada servik uterus. Data sekunder kedua adalah data *General Social Survey* 1982, yaitu data sampel tentang sikap warga Amerika Serikat terhadap suatu survei sosial yang diperoleh dari paket program poLCA versi 1.4 pada program R versi 3.02. Data survei sosial ini merupakan data sampel yang terdiri dari 4 variabel manifes yang digunakan sebagai indikator untuk menentukan sikap responden.

Kedua data sekunder ini dianalisis dengan menggunakan paket program poLCA versi 1.4 pada program R versi 3.0.2 (Anonim, 2013) dengan langkah-langkah sebagai berikut.

1. Melakukan pendugaan parameter dengan metode *Maximum likelihood Estimation* (MLE) yang dimaksimalkan dengan proses iterasi algoritma *Expectation-Maximum* (EM).
2. Memilih model terbaik berdasarkan nilai *AIC* dan nilai *BIC* terkecil.
3. Melakukan uji kriteria kecocokan model dengan membandingkan nilai G^2 dan χ^2 dengan nilai $\chi^2_{0.05;db}$.
4. Menginterpretasikan masing-masing kelompok yang telah dianalisis dengan analisis kelas laten.
5. Mengidentifikasi objek amatan yang akan masuk dalam masing-masing kelompok.

HASIL DAN PEMBAHASAN

Diagnosa Karsinoma

Sebagai studi kasus pertama dalam penggunaan analisis kelas laten adalah data sekunder *Diagnoses of carcinoma* (Karsinoma) yang diambil dari paket program poLCA versi 1.4 pada program R versi 3.02 dan Agresti (2002). Karsinoma adalah data sampel yang terdiri dari 118 objek amatan dengan 7 variabel manifes yang mewakili peringkat patalogi untuk mengetahui keadaan ada atau tidak ada penyakit karsinoma pada servik uterus. Ketujuh variabel patologi diberi label A, B, C, D, E, F, G dan setiap variabel terdiri dari dua kategorik, yakni kategorik “1” (Pr1) menyatakan tidak ada penyakit karsinoma pada servik uterus dan kategorik “2” (Pr2) menyatakan ada penyakit karsinoma pada servik uterus (Tabel 1).

Tabel 1. Data karsinoma dengan jumlah sampel 118 responden pada data karsinoma.

Variabel Manifes Patalogi							Frekuensi
A	B	C	D	E	F	G	
1	1	1	1	1	1	1	34
1	1	1	1	2	1	1	2
1	2	1	1	1	1	1	6
1	2	1	1	1	1	2	1
1	2	1	1	2	1	1	4
1	2	1	1	2	1	2	5
2	1	1	1	1	1	1	2
2	1	2	1	2	1	2	1
2	2	1	1	1	1	1	2
2	2	1	1	1	1	2	1
2	2	1	1	2	1	1	2
2	2	1	1	2	1	2	7
2	2	1	1	2	2	2	1
2	2	1	2	1	1	2	1
2	2	1	2	2	1	2	2
2	2	1	2	2	2	2	3
2	2	2	1	2	1	2	13
2	2	2	1	2	2	2	5
2	2	2	2	2	1	2	10
2	2	2	2	2	2	2	16
Jumlah Data							118

Dalam kasus data karsinoma ini, kelompok responden ada atau tidak adanya penyakit karsinoma berperan sebagai variabel laten, karena tidak dapat ditentukan secara langsung apakah objek amatan masuk ke dalam kelompok ada atau tidak ada penyakit karsinoma. Analisis dilakukan untuk 3 model kelas laten, yakni 1) model dengan 2 kelas laten; 2) model dengan 3 kelas laten; dan 3) model dengan 4 kelas laten. Hasil perhitungan dengan menggunakan program R versi 3.0.2 dengan paket program poLCA versi 1.4 dan diperoleh hasil seperti pada Tabel 2.

Tabel 2. Informasi kriteria kecocokan model pada data Karsinoma

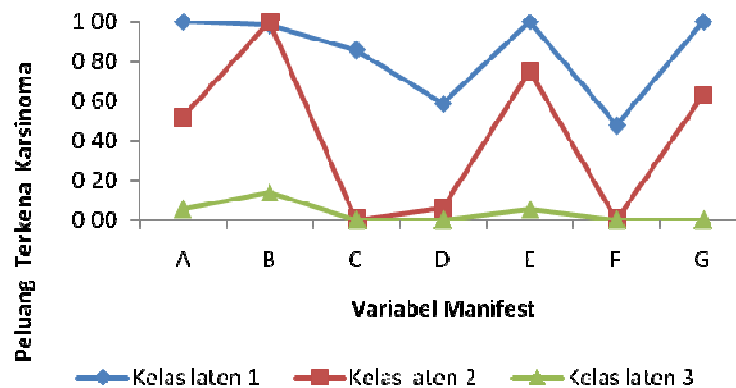
Model	2 kelas laten	3 kelas laten	4 kelas laten
Jumlah Pengamatan	118	118	118
Jumlah parameter yang diduga	15	23	31
Derajat bebas (db)	103	95	87
<i>maximum loglikelihood</i>	-317.2568	-293.705	-289.2858
<i>AIC</i>	664.5137	633.41	640.5717
<i>BIC</i>	706.0739	697.1357	726.4629
G^2 (ratio <i>Likelihood</i>)	62.36543	15.26171	6.423452
χ^2 (<i>goodness of fit</i> Chi-kuadrat)	92.64814	20.50335	10.08438
$\chi^2_{0.05;db}$	127.689	118.752	109.773

Dari Tabel 2 diperoleh informasi bahwa nilai AIC terkecil terdapat pada model tiga kelas laten (633,41), dan nilai BIC terkecil juga pada model tiga kelas laten (697,1357) maka model terbaik yang dipilih adalah model dengan tiga kelas laten. Untuk melihat apakah model tiga kelas laten telah memenuhi kriteria kecocokan model, maka dilakukan uji terhadap nilai G^2 dan χ^2 yang dibandingkan dengan nilai $\chi^2_{0.05;db}$. Oleh karena nilai $G^2 = 15.26717 < \chi^2_{0.05;db} = 118.752$ dan $\chi^2 = 20.50335 < \chi^2_{0.05;db} = 118.752$ maka model dengan tiga kelas laten memenuhi kriteria kecocokan model.

Selanjutnya dilakukan pendugaan parameter model untuk model tiga kelas laten. Pendugaan parameter ini bertujuan untuk mengetahui kelompok objek amatan. Ada tiga tipe objek amatan yang akan dikelompokkan, yakni kelompok negatif (tidak ada penyakit karsinoma), positif (ada penyakit karsinoma), ragu-ragu (tidak diketahui secara pasti apakah ada atau tidak ada penyakit karsinoma). Hasil pendugaan parameter untuk data Karsinoma ini disajikan pada Tabel 3 dan Gambar 1.

Tabel 3. Peluang bersyarat variabel manifes dengan model 3 kelas laten pada data karsinoma.

Variabel Manifes	Kategori	Kelas laten 1	Kelas laten 2	Kelas laten 3
A	Pr1	0.0000	0.4872	0.9427
	Pr2	1.0000	0.5128	0.0573
B	Pr1	0.0191	0.0000	0.8621
	Pr2	0.9809	1.0000	0.1379
C	Pr1	0.1425	1.0000	1.0000
	Pr2	0.8575	0.0000	0.0000
D	Pr1	0.4138	0.9424	1.0000
	Pr2	0.5862	0.0576	0.0000
E	Pr1	0.0000	0.2494	0.9449
	Pr2	1.0000	0.7506	0.0551
F	Pr1	0.5236	1.0000	1.0000
	Pr2	0.4764	0.0000	0.0000
G	Pr1	0.0000	0.3693	1.0000
	Pr2	1.0000	0.6307	0.0000






Gambar 1. Plot peluang bersyarat untuk variabel manifes dengan model tiga kelas laten pada data karsinoma

Pada kelas laten 1 diperoleh hasil bahwa peluang bersyarat untuk variabel manifes B positif terkena penyakit karsinoma adalah 0.9809, peluang bersyarat untuk variabel manifes C positif terkena penyakit karsinoma adalah 0.8575, peluang bersyarat untuk variabel manifes D positif terkena penyakit karsinoma adalah 0.5862, sedangkan peluang bersyarat untuk variabel manifes A, E, G positif terkena penyakit karsinoma adalah 1.0000. Hal ini menunjukkan bahwa kelas laten 1 dikategorikan sebagai kelompok kelas laten positif terkena penyakit karsinoma.

Pada kelas laten 2 diperoleh hasil bahwa variabel manifes A, B, E, G menyatakan positif terkena penyakit karsinoma sedangkan variabel manifes C, D, F menyatakan tidak terkena penyakit karsinoma (negatif). Hal ini menunjukkan bahwa kelas laten 2 dikategorikan sebagai kelompok kelas laten yang ragu-ragu apakah terkena penyakit karsinoma atau tidak.

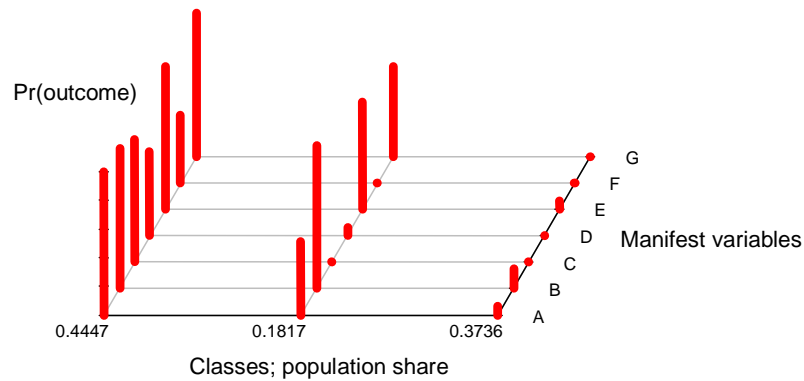
Demikian juga pada kelas laten 3, peluang bersyarat untuk variabel manifes A bebas dari penyakit karsinoma adalah 0,9427, peluang bersyarat untuk variabel manifes B bebas dari penyakit karsinoma adalah 0,8621, peluang bersyarat untuk variabel manifes E bebas dari penyakit karsinoma adalah 0,9449, Sedangkan untuk variabel manifes C, D, F, G, peluang bersyarat bebas dari penyakit karsinoma 1.0000. Hal ini mengindikasikan kelompok pada kelas laten 3 dikategorikan sebagai kelompok bebas dari penyakit karsinoma (negatif).

Dari Gambar 1 juga jelas terlihat bahwa untuk  Kelas laten 1 menunjukkan bahwa semua variabel manifes mempunyai peluang terkena penyakit Karsinoma di atas 0.50, yang artinya bahwa kelas laten 1 merupakan kelompok yang memiliki penyakit Karsinoma. Untuk  Kelas laten 2 menunjukkan ada 4 variabel manifes mempunyai peluang di atas 0,50 dan 3 variabel manifes mempunyai peluang di bawah 0.50 yang artinya bahwa kelas laten 2 merupakan kelompok yang ragu-ragu terhadap pengelompokan penyakit Karsinoma. Sementara untuk  Kelas laten 3 menunjukkan bahwa semua variabel manifes mempunyai peluang dibawah 0.50 yang artinya bahwa kelas laten 1 merupakan kelompok yang tidak memiliki penyakit Karsinoma.

Selanjutnya, analisis dilakukan untuk mengidentifikasi objek amatan yang masuk dalam masing-masing kelompok. Dari Tabel 4 dan Gambar 2 memperlihatkan bahwa 44,47% (53 objek amatan) masuk dalam kelompok yang secara konsisten dinilai positif mengandung penyakit Karsinoma, 37,36% (44 objek amatan) yang secara konsisten dinilai tidak mengandung penyakit Karsinoma, dan 18,17% (21 amatan) dinilai meragukan.

Tabel 4. Peluang objek amatan masuk dalam kelas laten pada data karsinoma.

Kelompok kelas laten	Peluang objek amatan masuk dalam kelas laten
Kelas laten 1 (Positif)	0.4447
Kelas laten 2 (Ragu-ragu)	0.1817
Kelas laten 3 (Negatif)	0.3736



Gambar 2. Pendugaan model kelas laten dengan model tiga laten kelas pada data karsinoma.

Survei Sosial

Sebagai ilustrasi kedua penggunaan analisis kelas laten adalah data sekunder *General Social Survey* 1982 (gss82) yang diambil dari paket program polCA versi 1.4 pada program R versi 3.02. Data survei sosial ini adalah data sampel tentang sikap warga Amerika Serikat terhadap sebuah survei sosial. Jumlah responden yang dilakukan wawancara ada sebanyak 1202 responden. Data survei terdiri dari 4 variabel manifes yang digunakan sebagai indikator untuk menentukan sikap responden dan disajikan pada Tabel 5.

Tabel 5. Statistik deskriptif variabel manifes data survei sosial

Variabel manifes	Indikator jawaban responden	Responden	Persentasi
Tujuan: Tujuan dilakukannya survei	Baik	919	76,46%
	Percaya	104	8,65%
	Pemborosan	179	14,89%
	Tidak menjawab	1	0,08%
Akurasi: Keakuratan survei	Sebagian besar tepat	625	52,00%
	Tidak tepat	577	48,00%
	Tidak menjawab	1	0,08%
Paham: Pemahaman responden tentang pertanyaan-pertanyaan pada survei	Paham	980	81,53%
	Tidak paham	222	18,47%
	Tidak menjawab	1	0,08%
Kooperatif Sikap responden saat dilakukan wawancara	Tertarik	1008	83,86%
	Kooperatif	159	13,23%
	Tidak sabar	35	2,91%

Dalam kasus survei sosial ini, tipe responden berperan sebagai variabel laten karena tipe responden tidak dapat diperoleh langsung tetapi dapat diperoleh dari variabel manifest, yaitu melalui sikap responden terhadap survei sosial tersebut. Untuk data survei sosial ini dilakukan analisis untuk 3 model kelas laten, yakni 1) model dengan 2 kelas laten; 2) model dengan 3 kelas laten; dan 3) model dengan 4 kelas laten. Hasil perhitungan dengan menggunakan program R versi 3.0.2 dengan paket program poLCA versi 1.4 diperoleh hasil seperti pada Tabel 6.

Tabel 6. Informasi kriteria dan kriteria kecocokan model pada data survei sosial

Model	2 kelas laten	3 kelas laten	4 kelas laten
Jumlah Pengamatan	1202	1202	1202
Jumlah parameter yang diduga	13	20	27
Derajat bebas (db)	22	15	8
maksimum <i>loglikelihood</i>	-2783.268	-2754.545	-2746.621
<i>AIC</i>	5592.536	5549.091	5547.242
<i>BIC</i>	5658.729	5650.926	5684.719
G^2 (ratio <i>Likelihood</i>)	79.33723	21.89203	6.042991
χ^2 (<i>goodness of fit</i> Chi-kuadrat)	93.25364	23.53222	5.114501
$\chi^2_{0.05;db}$	33.924	24.996	15.507

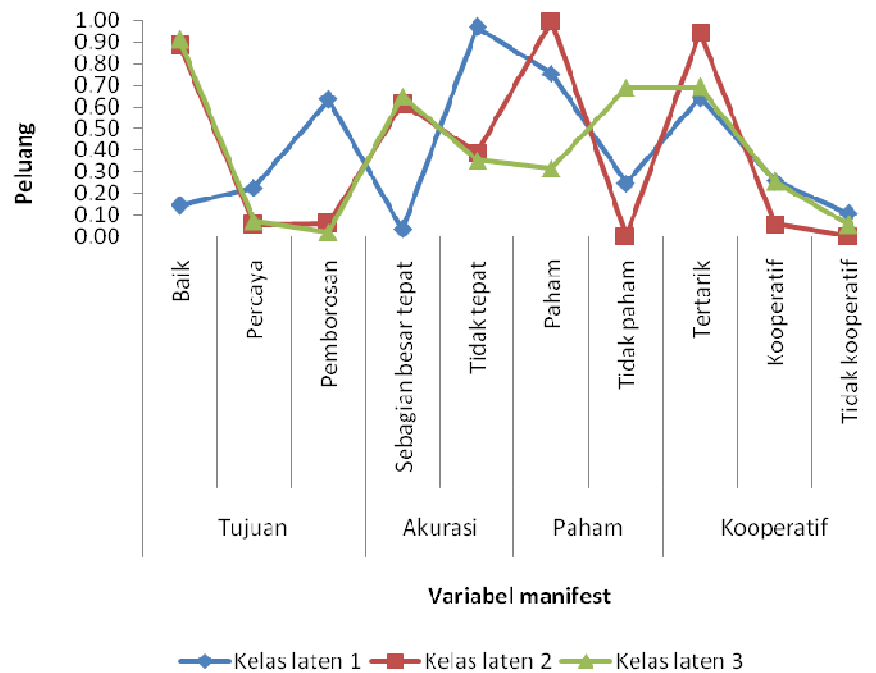
Dari Tabel 6 diperoleh informasi bahwa nilai *AIC* terkecil terdapat pada model dengan 4 kelas laten (5547.242), sedangkan nilai *BIC* terkecil terdapat pada model 3 kelas laten (**5650.926**). Dengan memperhatikan prinsip *parsimony*, jumlah parameter yang diduga pada model 3 kelas laten lebih kecil dari model 4 kelas laten dan juga jumlah data yang di analisis cukup besar maka model terbaik yang dipilih adalah model dengan 3 kelas laten.

Untuk melihat apakah model 3 kelas laten telah memenuhi kriteria kecocokan model, nilai G^2 dan χ^2 dibandingkan dengan nilai $\chi^2_{0.05;db}$. Oleh karena nilai $G^2 = 21.89203$ dan $\chi^2 = 23.53222$ lebih kecil dari nilai $\chi^2_{0.05;db} = 24.996$ maka model dengan tiga kelas laten memenuhi kriteria kecocokan model.

Selanjutnya dilakukan pendugaan parameter model untuk model 3 kelas laten, Pendugaan parameter ini bertujuan untuk mengetahui tipe ketiga kelompok objek amatan. Hasil pendugaan parameter untuk data survei sosial ini disajikan pada Tabel 7 dan Gambar 3.

Tabel 7. Peluang bersyarat variabel manifes dengan model 3 kelas laten pada data survei sosial

Variabel manifes	Kategori	Kelas laten 1	Kelas laten 2	Kelas laten 3
Tujuan	Baik	0.1427	0.8881	0.9117
	Percaya	0.2246	0.0532	0.0716
	Pemborosan	0.6327	0.0587	0.0167
Akurasi	Sebagian besar tepat	0.0313	0.6130	0.6478
	Tidak tepat	0.9687	0.3870	0.3522
Paham	Paham	0.7531	1.0000	0.3131
	Tidak paham	0.2469	0.0000	0.6869
Kooperatif	Tertarik	0.6410	0.9431	0.6897
	Kooperatif	0.2561	0.0569	0.2553
	Tidak sabar	0.1030	0.0000	0.0550






Gambar 3. Plot peluang bersyarat untuk variabel manifes dengan model tiga kelas laten pada data survei sosial

Pada kelas laten 1 diperoleh hasil bahwa variabel manifes “Tujuan” merepresenstasikan bahwa survei sosial tersebut adalah kategori “Pemborosan” sebesar 63.27%, variabel manifes “Akurasi” merepresenstasikan kategori “Tidak tepat” sebesar 96.87%, variabel manifes “Paham” merepresenstasikan kategori “Paham” sebesar 75.31%, dan variabel manifes “Kooperatif” merepresenstasikan kategori “Tertarik” sebesar 64.10%. Hal ini menunjukkan bahwa penilaian responden pada kelas laten 1 adalah tujuan dilakukan survei sosial merupakan pemborosan dan tidak tepat sasaran, namun responden memahami survei tersebut dan sangat kooperatif saat dilakukan wawancara. Berdasarkan ciri-ciri yang dimiliki, responden dalam kelas laten 1 masuk dalam kelompok dengan tipe responden “Skeptis”.

Pada kelas laten 2 diperoleh hasil bahwa variabel manifes "Tujuan" merepresentasikan kategori "Baik" sebesar 88.81%, variabel manifes "Akurasi" merepresentasikan kategori "Sebagian besar tepat" sebesar 61.30%, variabel manifes "Paham" merepresentasikan kategori "Paham" sebesar 100%, dan variabel manifes "Kooperatif" merepresentasikan kategori "Tertarik" sebesar 94.31%. Hal ini menunjukkan bahwa penilaian responden pada kelas laten 2 adalah tujuan dilakukan survei sosial sangat baik dan akurat, semua responden menyatakan bahwa survei yang dilakukan tepat sasaran dan responden sangat kooperatif saat dilakukan wawancara. Berdasarkan ciri-ciri yang dimiliki, responden dalam kelas laten 2 masuk dalam tipe responden "Optimis".

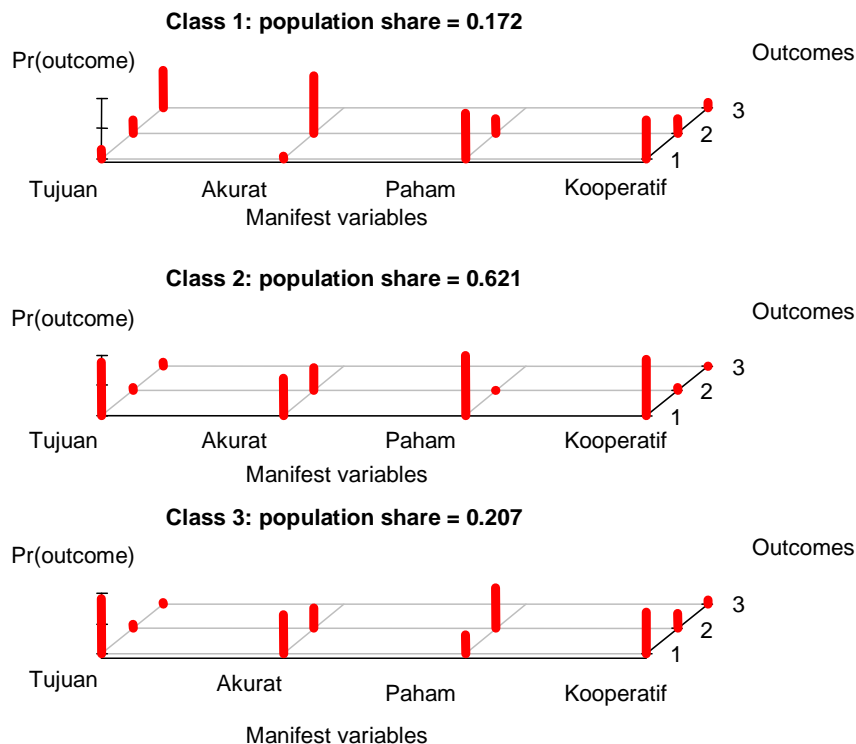
Demikian juga pada kelas laten 3, diperoleh hasil bahwa variabel manifes "Tujuan" merepresentasikan kategori "Baik" sebesar 91.17%, variabel manifes "Akurasi" merepresentasikan kategori "Sebagian besar tepat" sebesar 64.78%, variabel manifes "Paham" merepresentasikan kategori "Tidak paham" yaitu sebesar 68.69%, dan pada variabel manifes "Kooperatif" merepresentasikan kategori "Tertarik" sebesar 68,97%. Hal ini menunjukkan bahwa penilaian responden pada kelas laten 3 adalah tujuan dilakukan survei sosial sangat baik, akurat, dan sangat kooperatif saat dilakukan wawancara. Namun sebagian besar responden menyatakan bahwa survei yang dilakukan tidak paham. Berdasarkan ciri-ciri yang dimiliki, responden dalam kelas laten 3 masuk dalam tipe responden "Ideal".

Dari Gambar 1 juga jelas terlihat bahwa untuk  Kelas laten 1 menunjukkan bahwa responden cenderung memilih bahwa tujuan diadakan survei adalah pemborosan, tidak tepat sasaran, sebagian besar akurat, paham, dan tertarik. Untuk  Kelas laten 2 menunjukkan bahwa responden cenderung memilih bahwa survei diadakan untuk tujuan baik, tepat sasaran, paham, dan tertarik. Sementara untuk  Kelas laten 3 menunjukkan bahwa responden cenderung memilih bahwa survei diadakan untuk tujuan baik, tepat sasaran, dan tertarik, namun responden cenderung tidak paham tentang survei.

Selanjutnya, analisis dilakukan untuk mengidentifikasi objek amatan yang masuk dalam masing-masing kelompok. Dari Tabel 8 dan Gambar 4 memperlihatkan bahwa kelompok dengan tipe responden yang ideal diwakili 62,08% (746 responden) dari populasi, dengan yang lain 17,23% (207 responden) untuk tipe responden yang skeptis, dan 20,7% (249 responden) untuk tipe responden yang optimis.

Tabel 8. Peluang objek amatan masuk dalam kelas laten pada data survei sosial

	Kelas laten 1 (Skeptis)	Kelas laten 2 (ideal)	Kelas laten 3 (optimis)
Peluang objek amatan masuk dalam kelas laten	0.1723	0.6208	0.207



Gambar 4. Pendugaan model kelas laten dengan model tiga laten kelas pada data survei sosial

KESIMPULAN

Analisa kelas laten merupakan salah satu alat analisis untuk mengelompokkan data kategorik. Kedua studi kasus diasumsikan memenuhi kebebasan lokal. Hasil pengelompokan data kategorik pada sekunder diagnosa karsinoma dan survei sosial adalah sebagai berikut:

1. Diagnosa Karsinoma

Model terbaik adalah model dengan tiga kelas laten, yaitu kelompok yang secara konsisten dinilai positif ada penyakit karsinoma diwakili 44,47% (53 objek amatan), kelompok yang secara konsisten dinilai negatif terkena penyakit karsinoma diwakili 37,36% (44 objek amatan), dan 18,17% (21 objek amatan) dinilai masuk dalam kelompok meragukan apakah terkena penyakit karsinoma atau tidak.

2. Survei Sosial

Model terbaik adalah model dengan tiga kelas laten, yaitu kelompok dengan tipe responden ideal diwakili 62,08% (746 responden), kelompok dengan tipe responden optimis diwakili 20,7% (249 responden), dan kelompok dengan tipe responden skeptis diwakili 17,23% (207 responden).

DAFTAR PUSTAKA

- [Anonim]. (2013). Package “poLCA”. Diunduh tanggal 3 September 2013 dari cran.r-project.org/web/packages/poLCA/poLCA.pdf
- Agresti, Alan. 2002. *Categorical Data Analysis*, Second Edition. New Jersey, John Wiley & Sons, Inc.
- Collins, L. M. And S. T. Lanza. . 2010. *Latent Class and Latent Transition Analysis*, New Jersey, John Wiley & Sons, Inc
- Dean N., Raftery, A, R. (2010), Latent class analysis variable selection, *Ann Inst Stat Math*, Springer.
- Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, 6th Edition, New Jersey: Prentice-Hall.
- Linzer, Drew A. and Jerrey B. Lewis. 2011. poLCA: An R Package for Polytomous Variable Latent Class Analysis, *Journal of Statistical Software* . 42(10): 1-29.
- Moustaki, I. and Papageorgiou, I. (2004) Latent class models for mixed outcomes with applications in Archaeometry. *Computational Statistics and Data Analysis*, Vol. 48(3), 659-675.
- Magidson, J. & Vermunt, J.K. (2001). Latent Class Factor and Cluster Models, Bi-Plots, and Related Graphical Displays. *Sociological Methodology*, Vol. 31, 223–264.